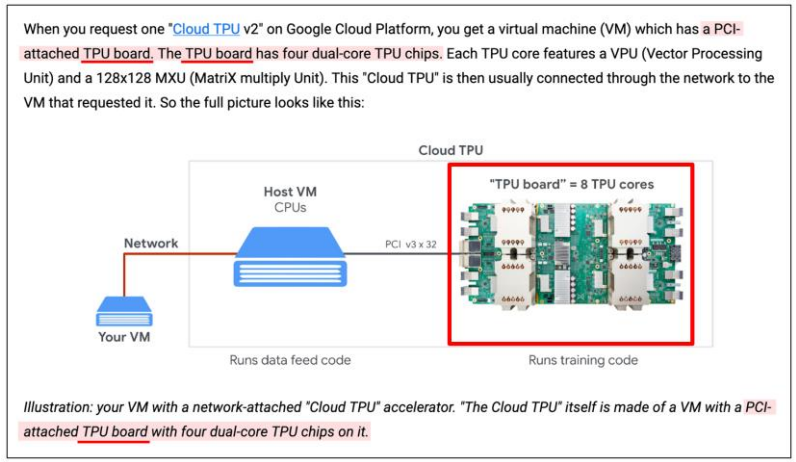
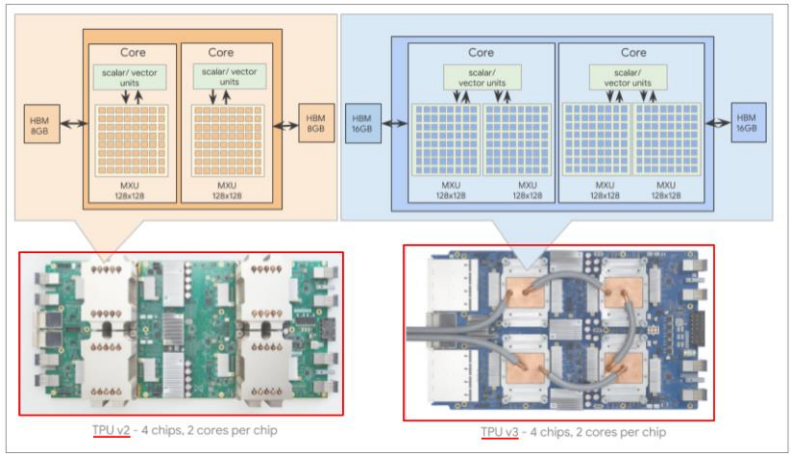
EXHIBIT A

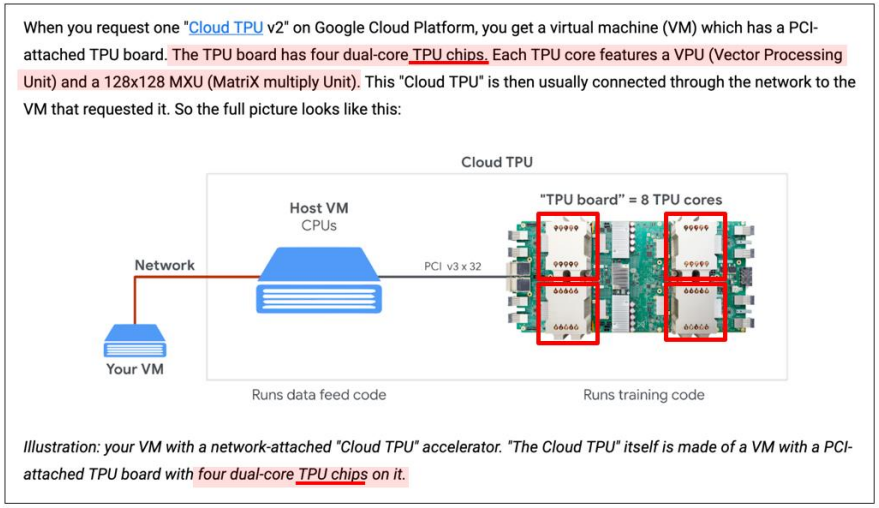
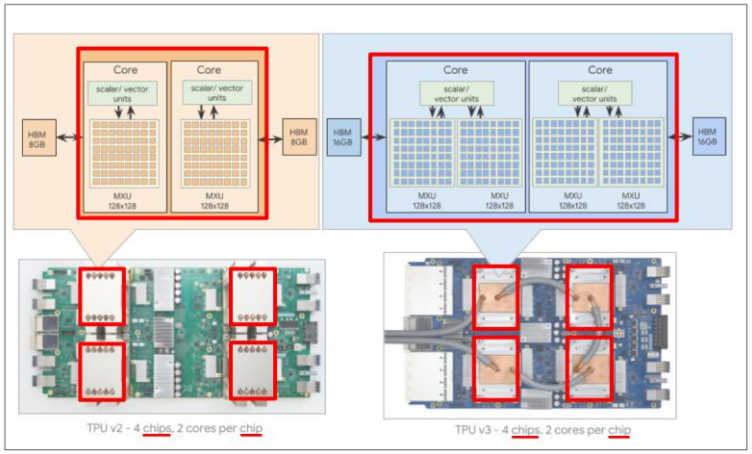
Exhibit A

U.S. Pat. No. 9,218,156

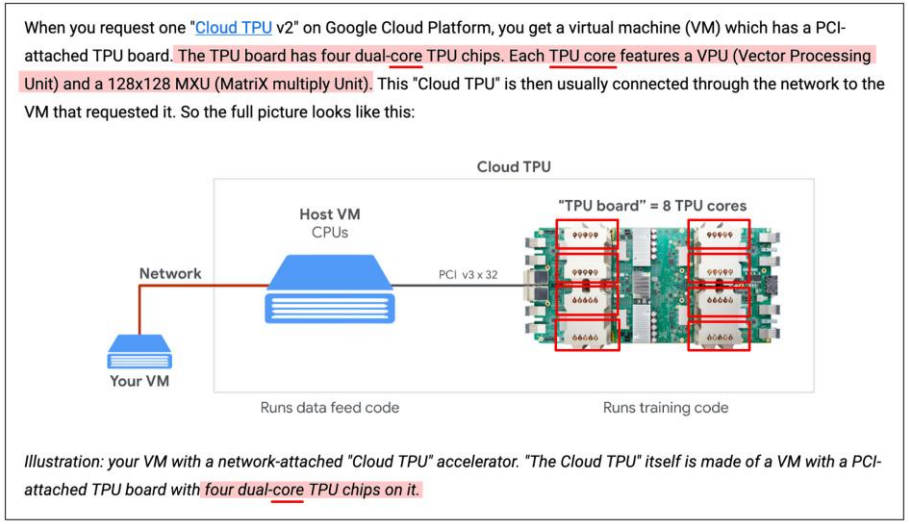
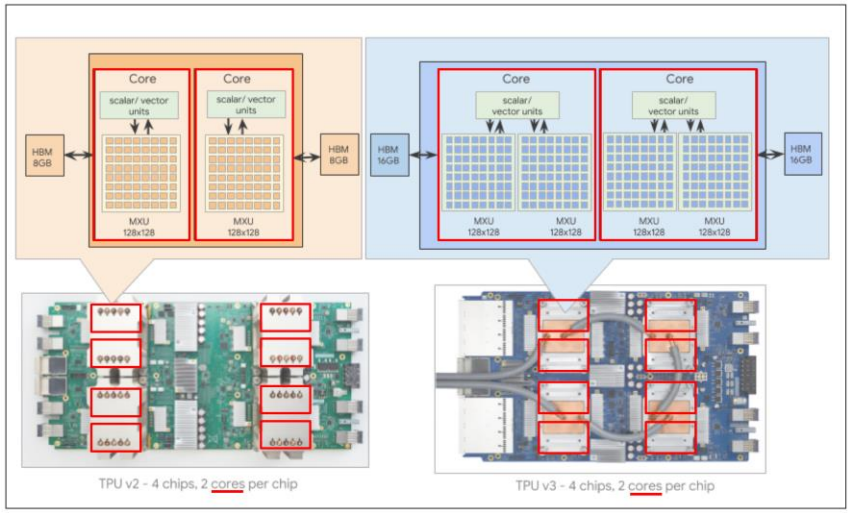
Claim 7

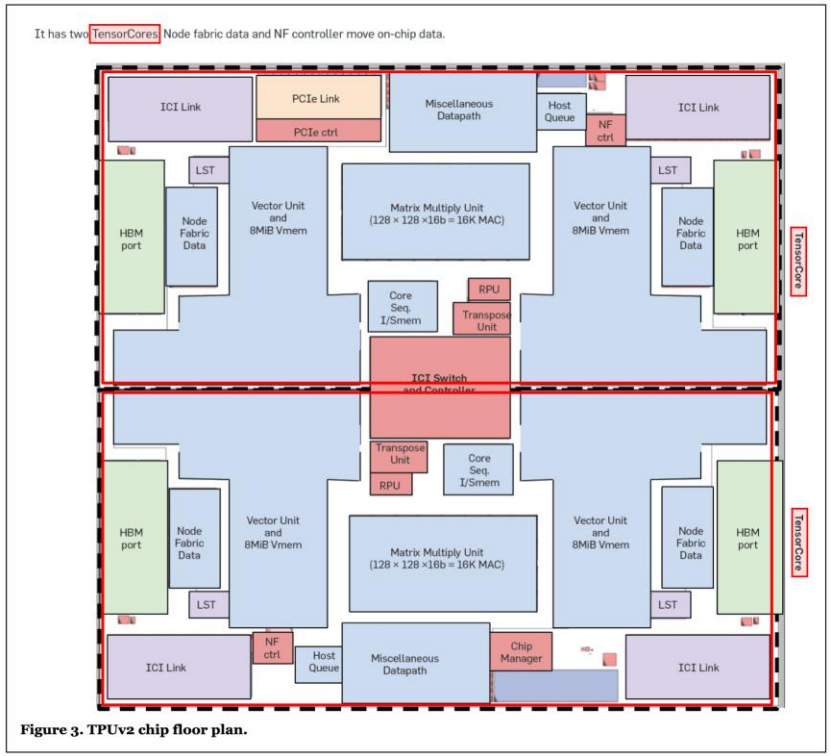
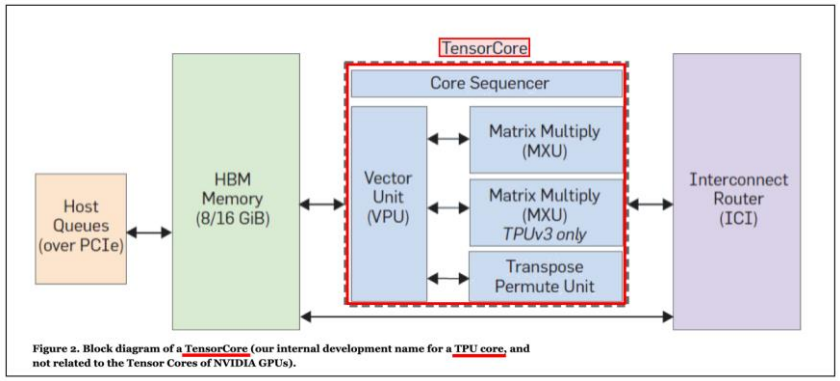
’156 PATENT	INFRINGEMENT EVIDENCE
7. A **device** comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	As demonstrated below, the Accused Products include multiple components that, separately and independently, meet all the requirements of the claimed “device.” For example, a “TPU Board” satisfies these requirements:  https://codelabs.developers.google.com/codelabs/keras-flowers-convnets/#2[1]  https://cloud.google.com/tpu/docs/system-architecture

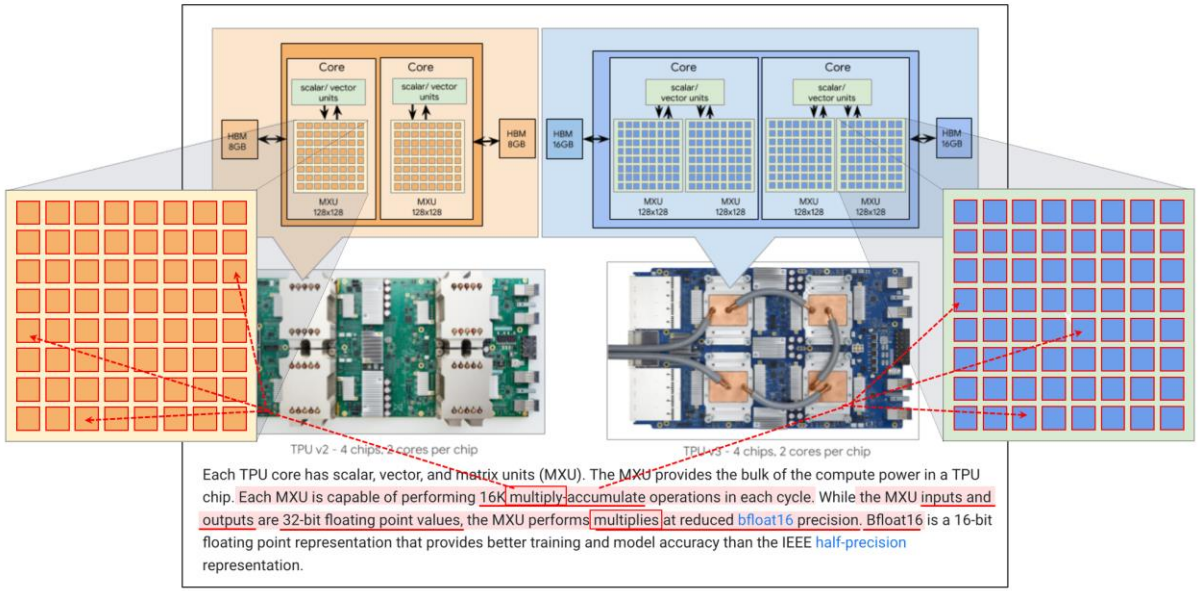
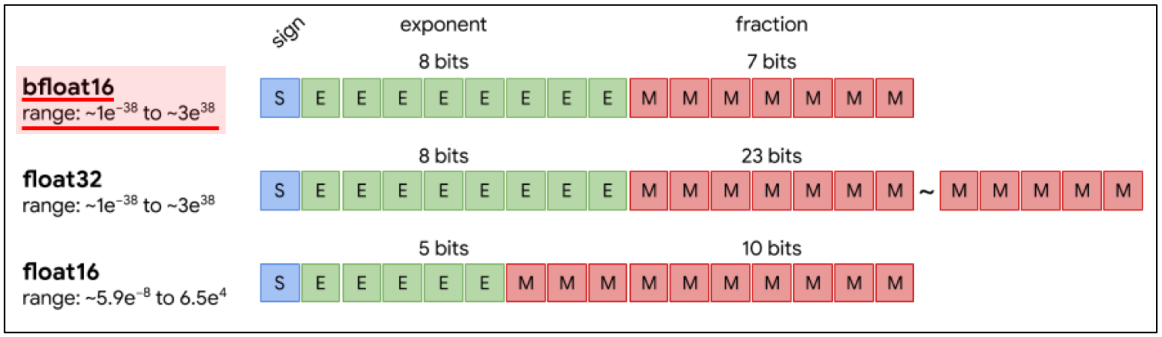
[1] Unless indicated otherwise, color-coded annotations have been added in order to identify relevant components and features of the Accused Products.

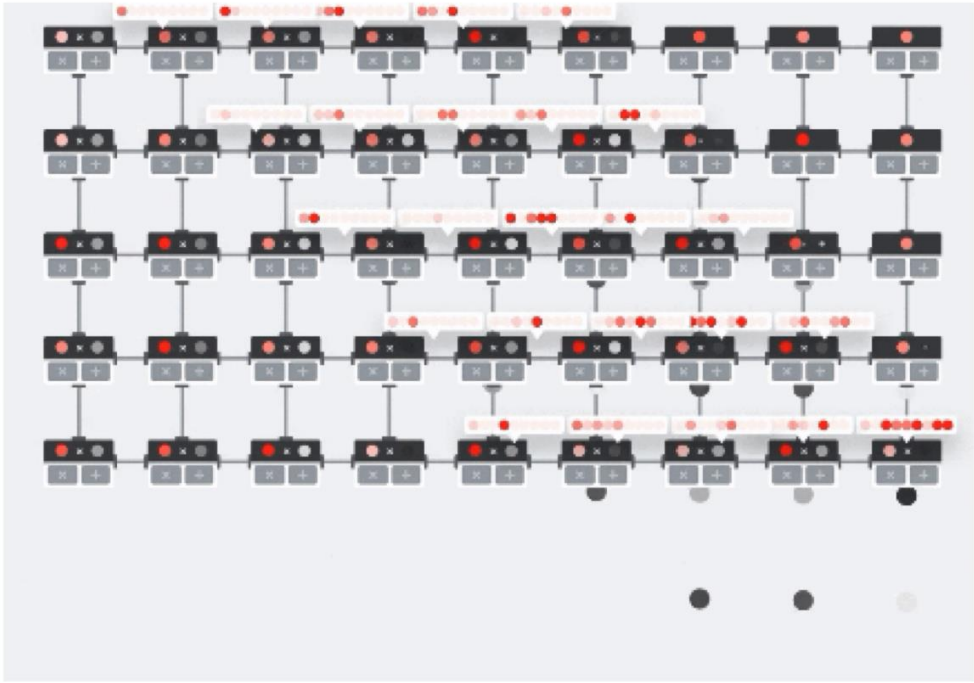
’156 PATENT	INFRINGEMENT EVIDENCE
7. A **device** comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	As demonstrated below, the Accused Products include multiple components that, separately and independently, meet all the requirements of the claimed “device.” For example, a “TPU Chip” satisfies these requirements:  https://codelabs.developers.google.com/codelabs/keras-flowers-convnets/#2[2]  https://cloud.google.com/tpu/docs/system-architecture *See also generally Norrie et al., “Google’s Training Chips Revealed: TPUv2 and TPUv3” (Presented at HotChips Conference, Aug. 2020)*

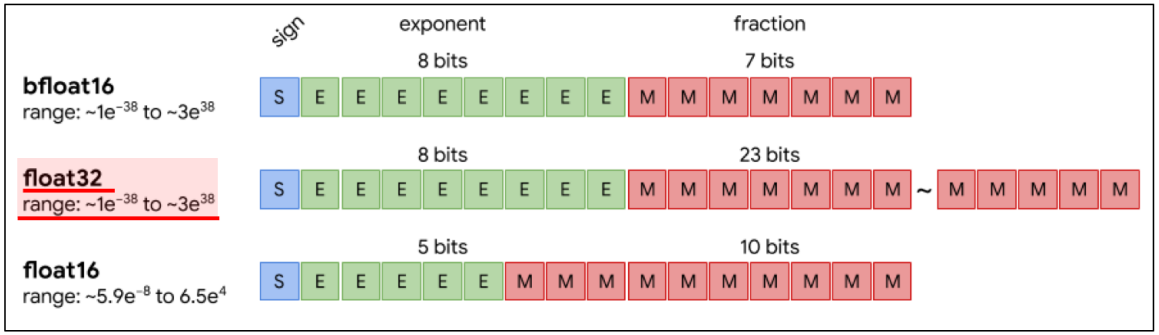
[2] Unless indicated otherwise, color-coded annotations have been added to the figures in this chart to highlight relevant teachings of the prior art.

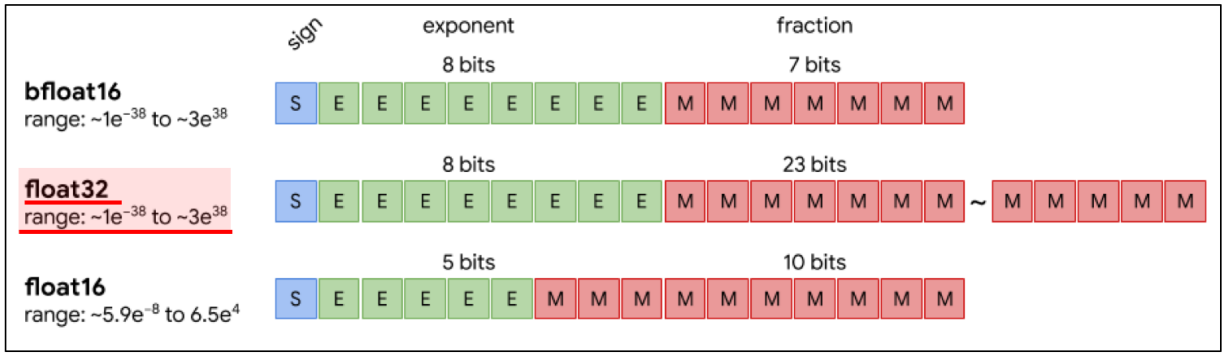
’156 PATENT	INFRINGEMENT EVIDENCE
7. A **device** comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	As demonstrated below, the Accused Products include multiple components that, separately and independently, meet all the requirements of the claimed “device.” For example, a “TPU Core” satisfies these requirements:  https://codelabs.developers.google.com/codelabs/keras-flowers-convnets/#2  https://cloud.google.com/tpu/docs/system-architecture

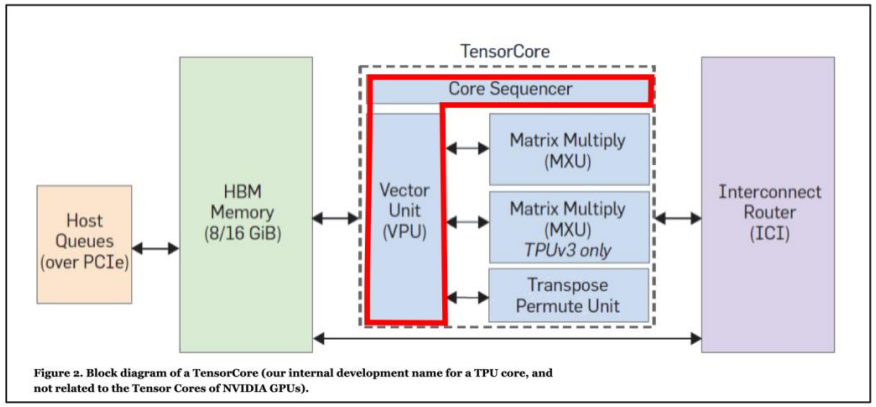
’156 PATENT	INFRINGEMENT EVIDENCE
7. A **device** comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	 **Figure 3. TPUv2 chip floor plan.** *https://cacm.acm.org/magazines/2020/7/245702-a-domain-specific-supercomputer-for-training-deep-neural-networks*  **Figure 2. Block diagram of a TensorCore (our internal development name for a TPU core, and not related to the Tensor Cores of NVIDIA GPUs).** *Id.*

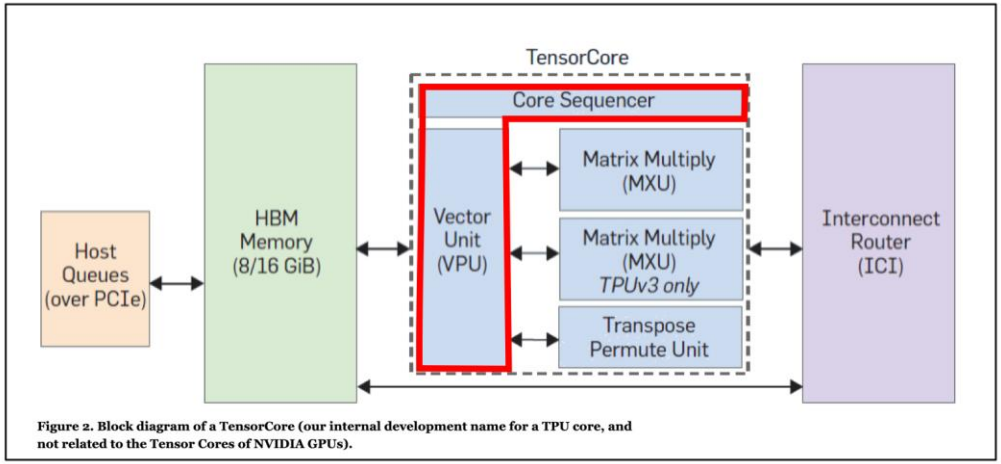
’156 PATENT	INFRINGEMENT EVIDENCE
7. A device comprising: **at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value,** wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	 *https://cloud.google.com/tpu/docs/system-architecture*  *https://cloud.google.com/tpu/docs/bfloat16*

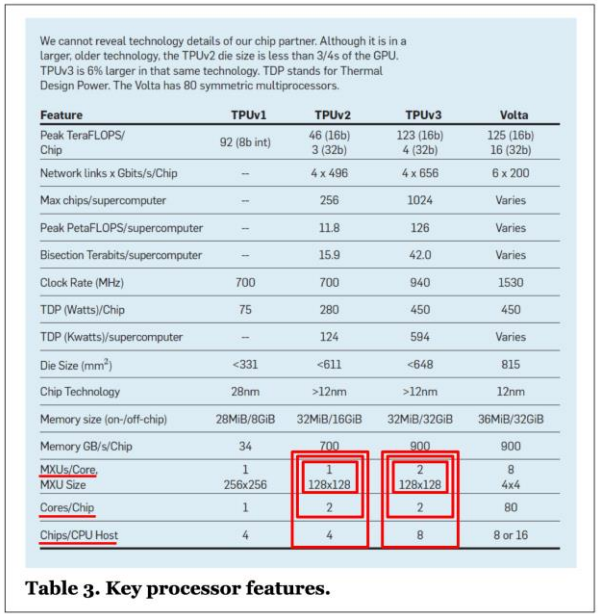
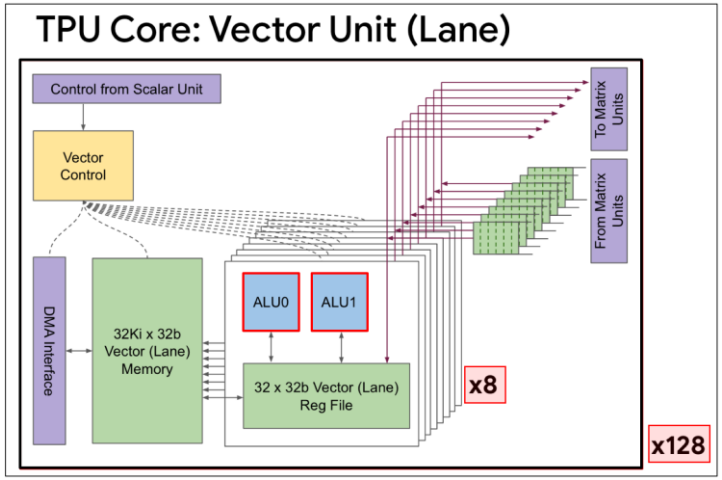
’156 PATENT	INFRINGEMENT EVIDENCE
7. A device comprising: **at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value,** wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	**Systolic array** The MXU implements matrix multiplications in hardware using a so-called "systolic array" architecture in which data elements flow through an array of hardware computation units. (In medicine, "systolic" refers to heart contractions and blood flow, here to the flow of data.) The basic element of a matrix multiplication is a dot product between a line from one matrix and a column from the other matrix (see illustration at the top of this section). For a matrix multiplication Y=X*W, one element of the result would be: `Y[2,0] = X[2,0]*W[0,0] + X[2,1]*W[1,0] + X[2,2]*W[2,0] + ... + X[2,n]*W[n,0]`  *Illustration: the MXU systolic array. The compute elements are multiply-accumulators. The values of one matrix are loaded into the array (red dots). Values of the other matrix flow through the array (grey dots). Vertical lines propagate the values up. Horizontal lines propagate partial sums. It is left as an exercise to the user to verify that as the data flows through the array, you get the result of the matrix multiplication coming out of the right side.* https://codelabs.developers.google.com/codelabs/keras-flowers-convnets/#2

’156 PATENT	INFRINGEMENT EVIDENCE
7. A device comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, **wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000** and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	• “Each TPU core has scalar, vector, and matrix units (MXU). The MXU provides the bulk of the compute power in a TPU chip. Each MXU is capable of performing 16K multiply-accumulate operations in each cycle. While **the MXU inputs and outputs are 32-bit floating point values**, the MXU performs multiplies at reduced bfloat16 precision. Bfloat16 is a 16-bit floating point representation that provides better training and model accuracy than the IEEE half-precision representation.” *https://cloud.google.com/tpu/docs/system-architecture* • “The following figure shows three floating-point[] formats • **fp32 - IEEE single-precision floating-point** • fp16 - IEEE half-precision floating point • bfloat16 - 16-bit *brain floating point*” *https://cloud.google.com/tpu/docs/bfloat16*  *Id.*

'156 PATENT	INFRINGEMENT EVIDENCE
7. A device comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and **for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input;** at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	• "Each TPU core has scalar, vector, and matrix units (MXU). The MXU provides the bulk of the compute power in a TPU chip. Each MXU is capable of performing 16K multiply-accumulate operations in each cycle. While **the MXU inputs and outputs are 32-bit floating point values**, the MXU performs multiplies at reduced bfloat16 precision. Bfloat16 is a 16-bit floating point representation that provides better training and model accuracy than the IEEE half-precision representation." *https://cloud.google.com/tpu/docs/system-architecture* • "The following figure shows three floating-point[] formats • **fp32 - IEEE single-precision floating-point** • fp16 - IEEE half-precision floating point • bfloat16 - 16-bit *brain floating point*" *https://cloud.google.com/tpu/docs/bfloat16*  *Id.* • "Because general-purpose processors such as CPUs and GPUs must provide good performance across a wide range of applications, they have evolved myriad sophisticated, performance-oriented mechanisms. As a side effect, the behavior of those processors can be difficult to predict, which makes it hard to guarantee a certain latency limit on neural network inference. In contrast, TPU design is strictly **minimal and deterministic** as it has to run **only one task at a time: neural network prediction**. You can see its simplicity in the floor plan of the TPU die." *https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu (emphasis in orig.)* • "In mathematics, computer science and physics, a **deterministic** system is a system in which **no randomness** is involved in the development of future states of the system. A **deterministic model will thus always produce the same output from a given starting condition or initial state**." *https://en.wikipedia.org/wiki/Deterministic_system* • For each of the possible valid inputs to the multiplication operation performed by the multipliers within the MXU, Singular has computed the result and compared it to the result of an exact mathematical calculation performed on the same inputs. The results of this test showed that for **more than 10%** of the possible valid inputs, the numerical value represented by the output signal of each MXU multiplier differs by **more than 0.2%** from the result of an exact mathematical calculation performed on the same inputs.

’156 PATENT	INFRINGEMENT EVIDENCE
7. A device comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; **at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit** wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	• “**Each of the cores on a TPU device can execute user computations (XLA ops)** independently.” *https://cloud.google.com/tpu/docs/system-architecture#pod* • “**TPUs use a VLIW architecture to express instruction-level parallelism to the many compute units of a TensorCore. XLA uses standard VLIW compilation techniques** including loop unrolling, instruction scheduling, and software pipelining to keep all compute units busy and to simultaneously move data through the memory hierarchy to feed them.” *https://cacm.acm.org/magazines/2020/7/245702-a-domain-specific-supercomputer-for-training-deep-neural-networks/fulltext* • “The ***Core Sequencer*** fetches *VLIW* (*Very Long Instruction Word*) instructions from the core's on-chip, software-managed Instruction Memory (Imem), executes scalar operations using a 4K 32-bit scalar data memory (Smem) and 32 32-bit scalar registers (Sregs), and forwards vector instructions to the VPU. **The 322-bit VLIW instruction can launch eight operations: two scalar, two vector ALU, vector load and store, and a pair of slots that queue data to and from the matrix multiply and transpose units. The XLA compiler schedules loading Imem via independent overlays of code, as unlike conventional CPUs, there is no instruction cache.**” *Id.* • “The ***Vector Processing Unit (VPU)*** performs vector operations using a large on-chip *vector memory* (*Vmem*) with 32K 128 x 32-bit elements (16MiB), and 32 2D *vector registers* (*Vregs*) that each contain 128 x 8 32-bit elements (4 KiB). **The VPU streams data to and from the MXU through decoupling FIFOs. The VPU collects and distributes data to Vmem** via *data-level parallelism* (2D matrix and vector functional units) and *instruction-level parallelism* (8 operations per instruction).” *Id.*  Figure 2. Block diagram of a TensorCore (our internal development name for a TPU core, and not related to the Tensor Cores of NVIDIA GPUs). *Id.*

’156 PATENT	INFRINGEMENT EVIDENCE
7. A device comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit; **wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine;** and, wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.	• “The ***Core Sequencer* fetches *VLIW* (*Very Long Instruction Word*) instructions from the core's on-chip, software-managed Instruction Memory (Imem), executes scalar operations using a 4K 32-bit scalar data memory (Smem) and 32 32-bit scalar registers (Sregs), and forwards vector instructions to the VPU**. The 322-bit VLIW instruction can launch eight operations: two scalar, two vector ALU, vector load and store, and a pair of slots that queue data to and from the matrix multiply and transpose units. The XLA compiler schedules loading Imem via independent overlays of code, as unlike conventional CPUs, there is no instruction cache.” *Id.* • “The ***Vector Processing Unit (VPU)*** performs vector operations using a large on-chip *vector memory* (*Vmem*) with 32K 128 x 32-bit elements (16MiB), and 32 2D *vector registers* (*Vregs*) that each contain 128 x 8 32-bit elements (4 KiB). **The VPU streams data to and from the MXU through decoupling FIFOs. The VPU collects and distributes data to Vmem via *data-level parallelism* (2D matrix and vector functional units) and *instruction-level parallelism* (8 operations per instruction).**” *Id.*  **Figure 2. Block diagram of a TensorCore (our internal development name for a TPU core, and not related to the Tensor Cores of NVIDIA GPUs).** *Id.* • “**Each of the cores on a TPU device can execute user computations (XLA ops)** independently.” *https://cloud.google.com/tpu/docs/system-architecture#pod* • “**TPUs use a VLIW architecture to express instruction-level parallelism to the many compute units of a TensorCore. XLA uses standard VLIW compilation techniques** including loop unrolling, instruction scheduling, and software pipelining to keep all compute units busy and to simultaneously move data through the memory hierarchy to feed them.” *https://cacm.acm.org/magazines/2020/7/245702-a-domain-specific-supercomputer-for-training-deep-neural-networks/fulltext*

’156 PATENT	INFRINGEMENT EVIDENCE
7. A **device** comprising: at least one first low precision high-dynamic range (LPHDR) execution unit adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000 and for at least X=5% of the possible valid inputs to the first operation, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input; at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit; wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine; and, **wherein the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.**	The Accused Products independently meet this claim limitation for each “device” identified above:  We cannot reveal technology details of our chip partner. Although it is in a larger, older technology, the TPUv2 die size is less than 3/4s of the GPU. TPUv3 is 6% larger in that same technology. TDP stands for Thermal Design Power. The Volta has 80 symmetric multiprocessors. **Table 3. Key processor features.** https://cacm.acm.org/magazines/2020/7/245702-a-domain-specific-supercomputer-for-training-deep-neural-networks/fulltext  TPU Core: Vector Unit (Lane) Control from Scalar Unit Vector Control To Matrix Units From Matrix Units DMA Interface 32Ki x 32b Vector (Lane) Memory ALU0 ALU1 32 x 32b Vector (Lane) Reg File x8 x128 *Norrie et al., “Google’s Training Chips Revealed: TPUv2 and TPUv3” (Presented at HotChips Conference, Aug. 2020)*

Feature	TPUv1	TPUv2	TPUv3	Volta
Peak TeraFLOPS/ Chip	92 (8b int)	46 (16b) 3 (32b)	123 (16b) 4 (32b)	125 (16b) 16 (32b)
Network links x Gbits/s/Chip	--	4 x 496	4 x 656	6 x 200
Max chips/supercomputer	--	256	1024	Varies
Peak PetaFLOPS/supercomputer	--	11.8	126	Varies
Bisection Terabits/supercomputer	--	15.9	42.0	Varies
Clock Rate (MHz)	700	700	940	1530
TDP (Watts)/Chip	75	280	450	450
TDP (Kwatts)/supercomputer	--	124	594	Varies
Die Size (mm^2)	<331	<611	<648	815
Chip Technology	28nm	>12nm	>12nm	12nm
Memory size (on-/off-chip)	28MiB/8GiB	32MiB/16GiB	32MiB/32GiB	36MiB/32GiB
Memory GB/s/Chip	34	700	900	900
MXUs/Core, MXU Size	1 256x256	1 128x128	2 128x128	8 4x4
Cores/Chip	1	2	2	80
Chips/CPU Host	4	4	8	8 or 16